Information Dynamics in the Networked World

Shiwu Zhang

Based on the paper by B.A. Huberman and L.A. Adamic from HP Labs

Three Studies on HP Email Network

- Develops an automated method applying a betweenness centrality algorithm to rapidly identify communities
- Analyzes email patterns to model information flow in social groups, taking into account the decay derived from the organization distance
- Simulates Milgram's small world experiment on the HP labs email network according to different search strategies

HP Labs Email Network

- A set of over one million email messages collected over a period of roughly two months at HP labs in Palo Alto, an organization of approximately 400 peoples
- The only pieces of information used from each email are the names of the sender and receiver
- Construct a graph based on email data, in which vertices represent people and edges are added between people who exchange at least a threshold number of email messages

Identifying Communities-Betweenness

- Definition: the number of all pair shortest paths that traverse an edge, it is firstly proposed by Freeman
- Distinguishes inter-community edges, which link many vertices in different communities and have high betweenness, from intra-community edges, whose betweenness is low

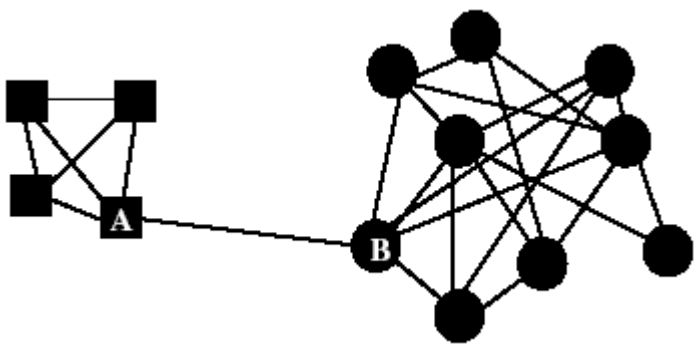
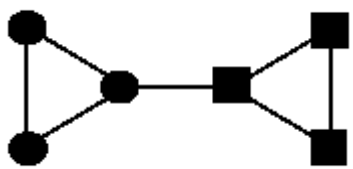
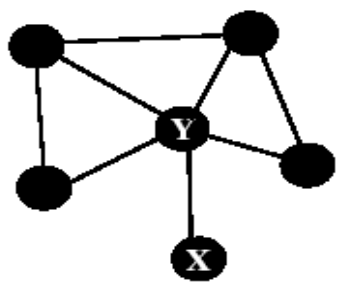


Figure 1

Identifying Communities-Removal of edges with highest betweenness

- Wilkinson's algorithm: repeatedly identifies inter-community edges of large betweenness and removes them
- Removal of an edge strongly affects the betweenness of many others
- Stopping criterion: for components of size>=6, the highest betweenness of any edge in the component be equal to or less than N-1





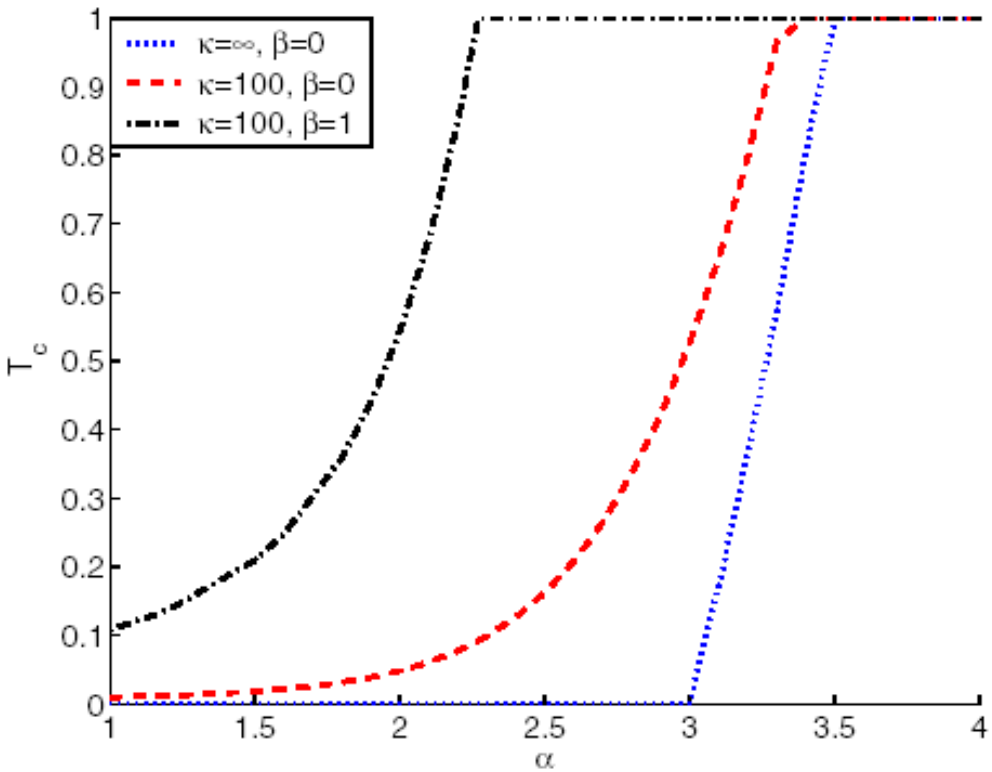
Identifying Communities- Randomness factor

- The algorithm has flaws
  - Too slow, we have to compute the betweenness after each removal
  - The order in which edges are removed affects the final structureAdd
- Introduce Randomness into the algorithm
  - M centers to all nodes
  - Repeated until the graph has been separated into communities
- Aggregation after applying the modified process n times
  - N community structures
  - compare/aggregation
  - A(50)B(50)C(50)D(50)E(25)F(5)

Identifying Communities - Results

- Graph constructed
  - 3 months, 485 employees
  - Add an edge when two people had exchanged at least 30 emails and at least 5 in both direction
  - 367 nodes with 1110 edges
- Identified communities
  - 60 additional distinct communities within the giant component
  - 49 of 66 communities consisted of individuals entirely in one lab
- Identifying leadership roles
  - A standard force-directed spring algorithm
  - It does not use any information about the actual organization structure, while high level managers are placed close to the center of the graph

Information Flow in Social Group - Numerical Results

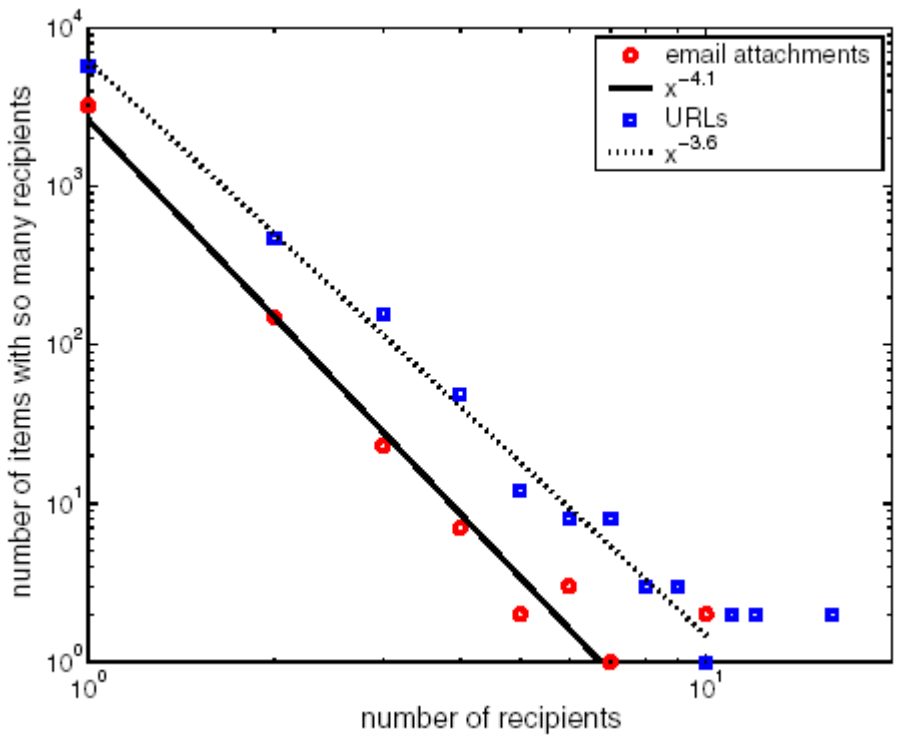


κ : the exponential cutoff, α: power

β: decay constant, T_c: transmissibility above which <s> will exceed 1%

- For α<3, epidemics encompassing more than 1% vertices occur for arbitrarily small *T*
- Adding a cutoff, a non-zero critical transmissibility was found, and for α=2 (real-world network), Tc is still near zero
- Imposing a decay in transmissibility, Tc rises substantially
- The information may not spread over the network so easily

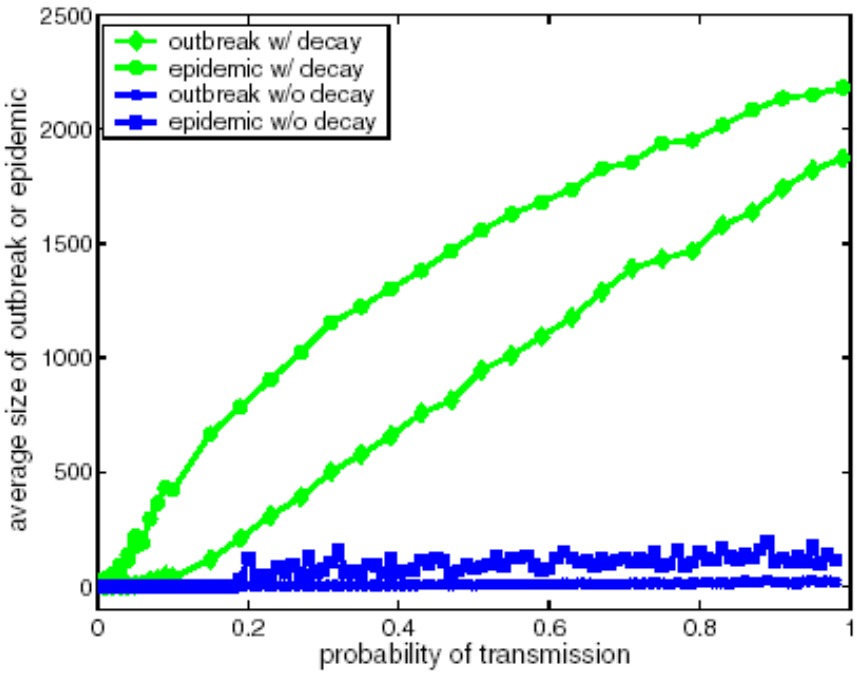
Information Flow in Social Group- Empirical validation



The network of people is limited, distinct from the spread of a virus. 40 individuals was gathered.

- The email data was restricted to include messages which had been forwarded at least one time
- The median number of the message was 2200.
- 3401 attachments, 6370 URLs,
- Only a small fraction reached more than 1 individuals, few reached more than 5

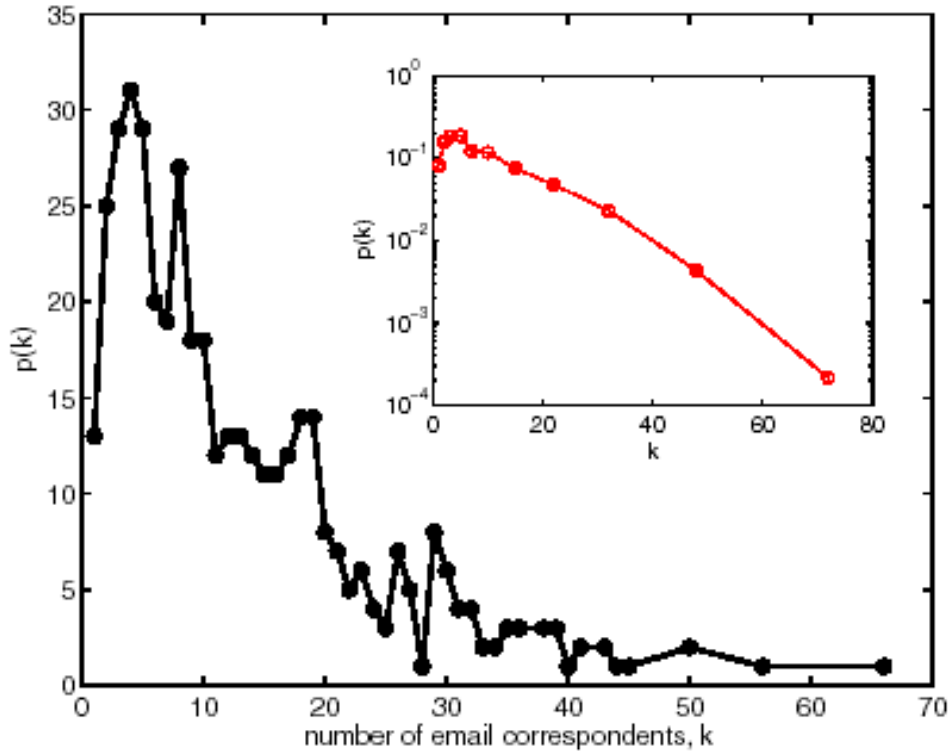
Information Flow in Social Group- Effect of Decay



The effect of decay in the transmission probability on the email graph. The email graph follows a power-law degree distribution

- Selecting a random initial sender to infect (following the email logs)
- The sender had a infecting probability p
- Imposing a decay in the organizational hierarchical distance d
- Without decay, the epidemic threshold falls below $p = 0.01$; with decay, the threshold is 0.2, the epidemic size is limited to about 50 individuals, even for $p = 1$

Small World Search- The network

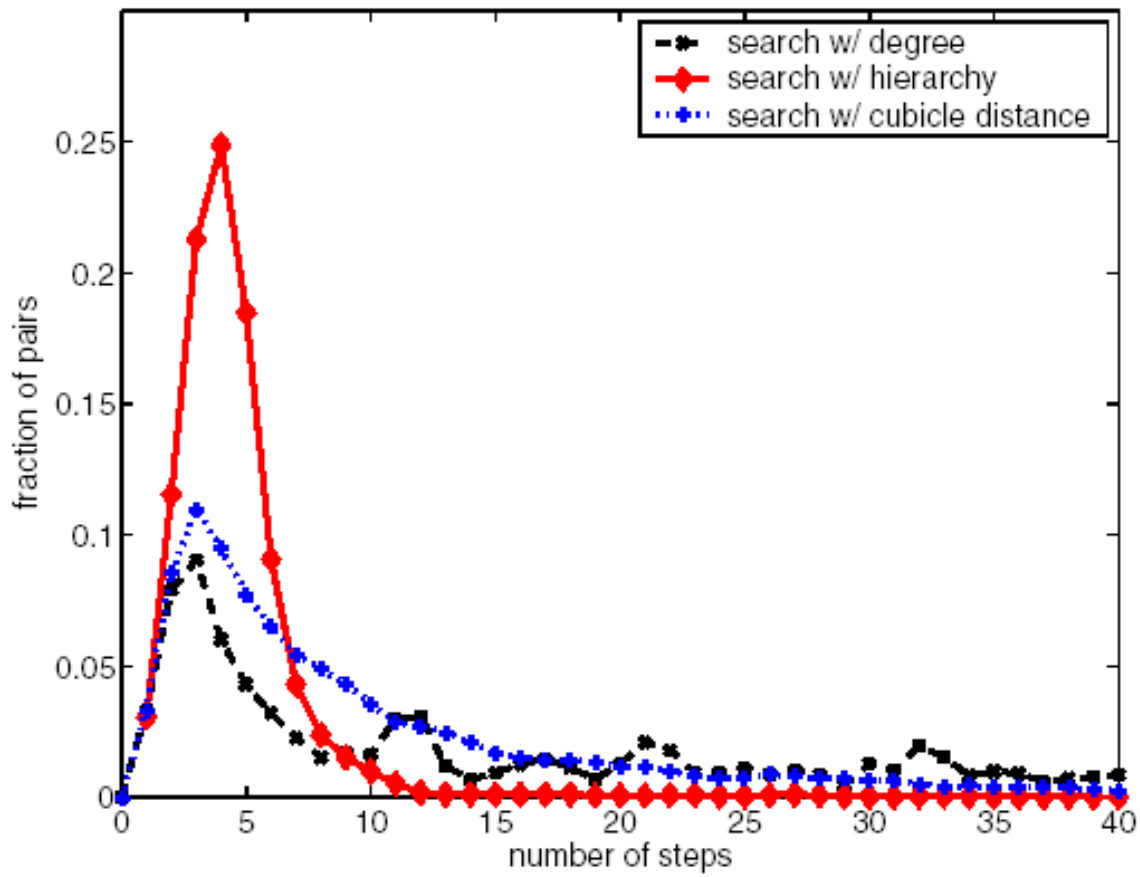


- A social contact was defined to be someone with whom an individual had exchanged at least 6 emails over 3 months

Small World Search- Three Strategies

- Best connected
  - Effective in power-law networks with exponents close to 2
  - Poor in a Poisson degree distribution that has an exponential tail
  - Median steps: 17, average steps: 40, which reflect the fact that some individuals who do not have many links and are not connected to highly connected individuals are difficult to locate using the strategy
- Closest to the target in the organizational hierarchy
  - Hierarchical distance
  - Median number: 4, mean: 4.7
- Sitting in closest physical proximity to the target
  - A geography information
  - Median number: 7, mean: 12

Small World Search- Results



Conclusion

- Betweenness
  - An effective method that can be used to identify many communities of other types
- Decay in transmissibility
  - The information spread is limited
- Search strategy
  - Hierarchical knowledge is beneficial to search in email network

References:

Information dynamics in the networked world, Huberman and Adamic

Information flow in social groups, F. wu, Huberman, et al.

A method for finding communities of related genes, Wilkinson and Huberman

How to search a social network, Adamic and Adar